

A Natural Law of Succession

Paper by Eric Sven Ristad,
Princeton University, May 1995

Presented by Ben Leong

CSE 254, May 16, 2002

Overview

Background

Motivation, problem framework

Previous methods – Laplace and Lidstone

String methods

Uniform subset or cardinality(natural)

Experiments

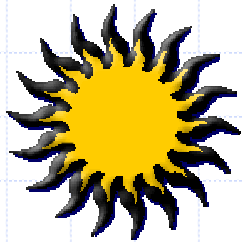
Sunrise, files, natural language

Will the sun rise tomorrow?

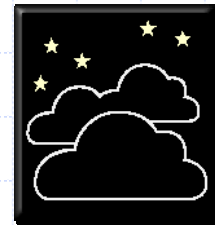
Question: What is the probability the sun will/will not rise tomorrow?

Data: The sun has risen each day in past 5,000 years (last 1,826,213 days)

Possible answer:



$$\frac{1,826,214}{1,826,215} \text{ vs. } \frac{1}{1,826,215}$$



Framework for whole paper

Notation:

1. Alphabet A , with k distinct symbols
2. An observed string x^n of length n over alphabet A
3. q distinct symbols in x^n
4. Exactly n_i occurrences of i^{th} symbol in x^n

$$q = \left| \{n_i \text{ s.t. } n_i > 0\} \right|$$

Task: Predict next symbol

Given:

Alphabet A with k distinct symbols

History of n examples

Count n_i of i^{th} symbol in last n examples in history

Estimate:

$$p(i \mid \{n_i\}, n)$$

Need to estimate $p(i|\{n_i\},n)$ well

For each symbol i , find the associated hidden parameter p_i that represents the probability of occurrence

“Arguably the single most important parameter estimation problem in statistics”

An accurate estimate of this is central to any problem that uses a discrete probability distribution.

- Sequence prediction with Markov models
- Classification with decision trees
- And more...

Standard methods

Two common solutions for this parameter estimation task:

- Laplace's law of succession
- Lidstones's law of succession

Laplace's law

Definition:

$$p_L(i|\{n_i\}, n) = \frac{n_i + 1}{n + k}$$

For $n = 0$,

$$p_L(i|\{n_i\}, n) = \frac{1}{k}$$

a uniform distribution

For $n_i = 0$,

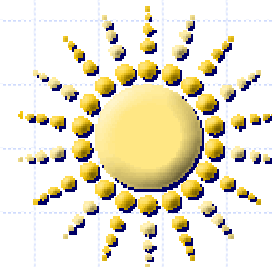
$$p_L(i|\{n_i=0\}, n) = \frac{1}{n + k}$$

Rising sun and Laplace's law

$n_0=0$ (failed sunrises)

$n_1=1,826,213$ (successful sunrises)

$k=2$

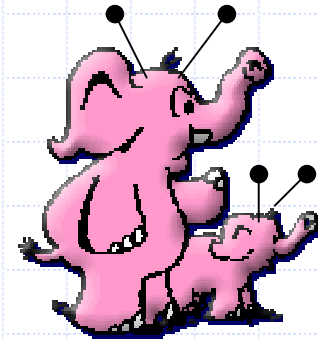


$$p(\text{no sun} \mid \{n_i\}, n) = \frac{1}{1,826,215}$$

Compared to New Jersey “Pick 6” and CA lotto:

$$\frac{1}{9,366,819} \quad \text{and} \quad \frac{1}{41,416,353}$$

Elephants on Mars



Question: How many elephants live on Mars?

Data: Mars has approximately 54,055,250 square miles. Assuming 1 square mile supports 1 elephant. The possible number of elephants on Mars is 0-54,055,250.

Each day we count the elephants on Mars.

We'd have the same confidence as the sun rising after 270,270,917 millennia of observing no elephants.

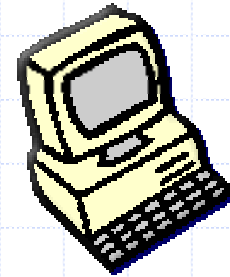
Random number generator?

Supposed to generate 32-bit random numbers.

After $2^{32}-1$ trials, $2^{32}-1$ distinct numbers are observed.

Laplace's Probability of observing the only unseen number:

$$\frac{0 + 1}{(2^{32} - 1) + 2^{32}} = \frac{1}{2^{33} - 1}$$



Lidstone's Law

$$p_\lambda(i|\{n_i\}, n) = (n_i + \lambda)/(n + k\lambda)$$

Can be considered a mixture of (n_i/n) and $(1/k)$

for $\mu = n/(n + k\lambda)$:

$$p_\lambda(i|\{n_i\}, n) = \mu(n_i/n) + (1-\mu)(1/k)$$

Example:

Given: prior $p(i) = \frac{1}{k}$ and data $\{n_i\}, n$

Then Bayesian posterior is: $p(i | \{n_i\}, n) = \frac{n_i + 1}{n + k}$

Problems with Lidstone's law

$\lambda < 1$ implies more trust in observed relative frequency. $\lambda > 1$ implies less trust in observations and more in the uniform distribution.

In the case of the sunrise example, prior knowledge tells us $P(\text{no sunrise}) = 0$. No value of λ expresses this.

We cannot set a good value of λ if the prior knowledge is incomplete or not easily describable.

New Approach: Priors on Strings

Instead of priors on multinomial distributions
We impose constraints s.t. simple strings are
more probable than complicated ones.

What's a "natural" string?

Answer:

Natural strings are a sequence of discrete symbols that only use a proper subset of alphabet A

Example 1:

Computer files with 8-bit bytes have 256 symbols but ASCII is limited to 95 symbols.

Example 2:

Average English sentences draw from less than 20,000 possible words, compared to 10^6+ English word terms in Webster's 3rd.

Alternative priors on strings

We consider 2 interpretations of “natural”:

- ◆ Each nonempty subset of the alphabet is equally likely
- ◆ Each nonzero cardinality is equally likely

In both cases, a string using a small subset of A is more likely than a string from a larger subset of A

Equal Subset of A Probability

Each subset of A gets an equal slice of probability space.

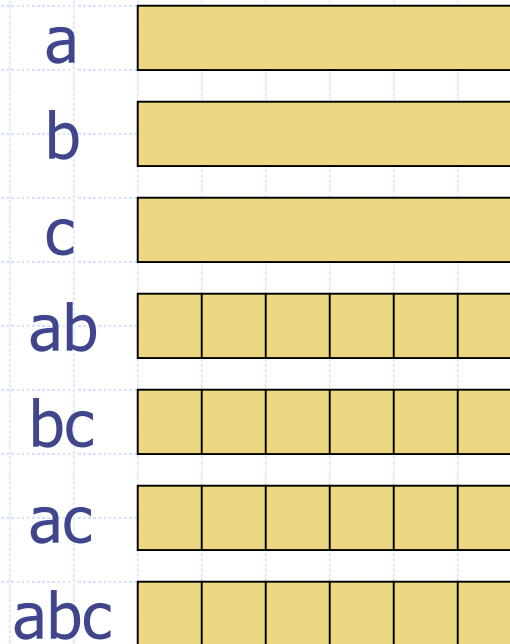
Each of these slices is then equally divided among all strings.

Example:

$$A = \{a,b,c\}$$

$$n = 3, k=3$$

Subset



Uniform prior on subsets

$$p_S(x^n | n) = \left(\left(\sum_{i=1}^{\min(k,n)} \binom{k}{i} \right) \binom{n-1}{q-1} \binom{n}{\{n_i\}} \right)^{-1}$$

1st term – uniform prior(u.p) of nonempty subsets

2nd term – u.p. of n_i , given subset of A

3rd term – u.p. of strings given $\{n_i\}$, n

Equal Cardinality Probability

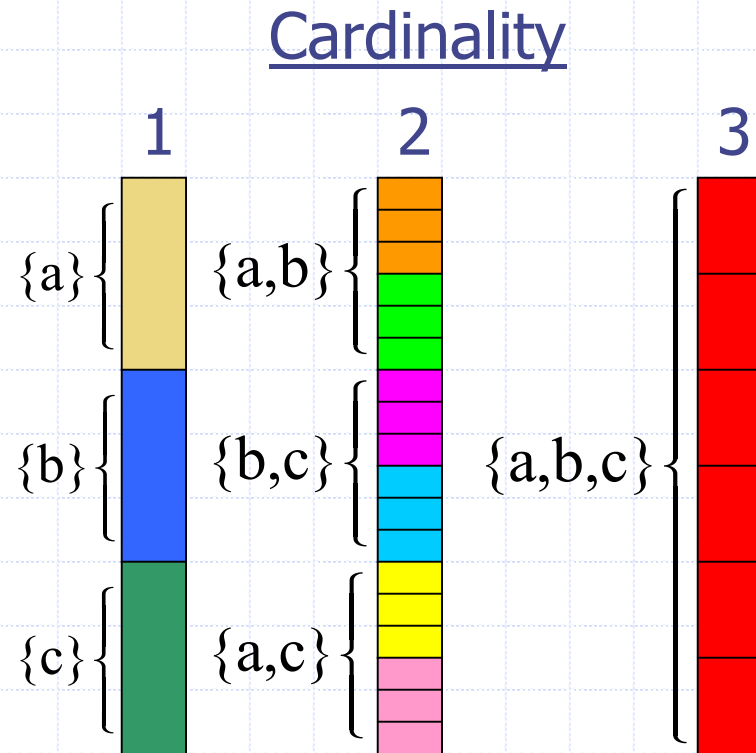
Each Cardinality gets an equal slice of probability space.

Each slice is then equally divided among all strings.

Example:

$$A = \{a,b,c\}$$

$$n = 3$$



Uniform Cardinality Prior

$$p_C(x^n | n) = \left(\min(k, n) \binom{k}{q} \binom{n-1}{q-1} \binom{n}{\{n_i\}} \right)^{-1}$$

1st term – uniform prior (u.p.) over cardinalities

2nd term – u.p. over subset of A with given size q

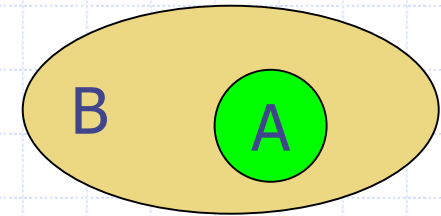
3rd term – u.p. over $\{n_i\}$ for the chosen subset of A

4th term – u.p. over strings for the chosen $\{n_i\}$ and n

First term explained

Recall:

If $A \rightarrow B$, then $P(A) = P(A \cap B)$



Also:

$$P(A \cap B) = P(A|B)P(B)$$

$$\begin{aligned} p(x^n | n) &= p(x^n \cap q | n) \\ &= p(x^n | q, n) \underbrace{p(q | n)}_1 \\ &= \frac{1}{\min(k, n)} \end{aligned}$$

Second term explained

$$p(x^n|n) = \overbrace{p(x^n|q,n)}^{\text{The rest}} \overbrace{p(q|n)}^{1^{\text{st}} \text{Term}}$$

$$\text{The rest} = p(x^n \cap \{i:n_i > 0\} | q, n) p(\underbrace{\{i:n_i > 0\}}_{\binom{k}{q}} | q, n)$$

Example:

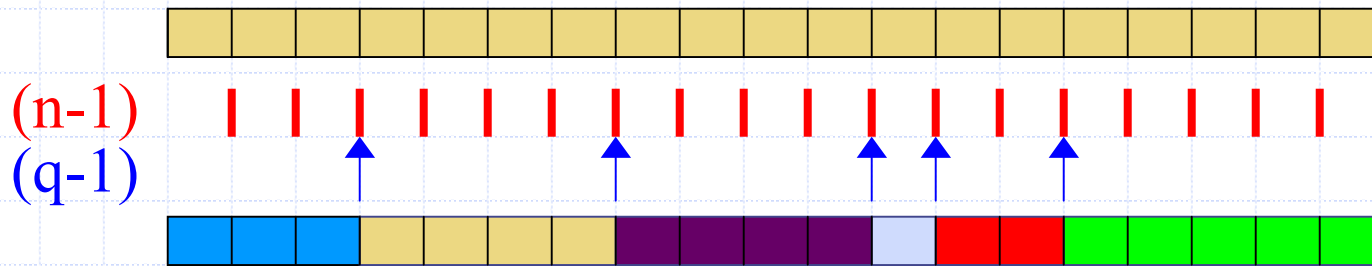
$$A = \{1, 2, 3, 4\}, q = 2, k = 4$$

Solutions:

$$\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}$$

Third term explained

Reasoning for 3rd term = $\binom{n-1}{q-1}$



There are $(n-1)$ positions between symbols.

We place $(q-1)$ dividers to determine $\{n_i\}$ for each $n_i > 0$.

(This example: $n=19, q=6$)

Fourth term explained

How many different ways can $\{n_i\}$ be arranged in a string of length n ?

$$\binom{n}{\{n_i\}} = \frac{n!}{n_1! n_2! n_3! \dots n_q!}$$

The succession probability of i

The probability that the symbol i will follow x^n relative to the probability of any symbol j in A following x^n .

$$p(i | x^n, n) \doteq \frac{p(x^n i | n + 1)}{\sum_{j=1}^k p(x^n j | n + 1)}$$

A New Law of Succession

Based on uniform subset prior:

$$p_S(i | \{n_i\}, n) =$$

$$\begin{cases} \frac{(n_i + 1)(n + 1 - q)}{((n + q)(n + 1 - q) + q(k - q))}, & \text{if } n_i > 0 \\ q / ((n + q)(n + 1 - q) + q(k - q)), & \text{otherwise} \end{cases}$$

Quick Check

Recall: Natural Strings leverage fact that $q < k$

What happens when $q = k$?

$$\begin{aligned} p_S(i | \{n_i\}, n) &= \frac{(n_i + 1)(n + 1 - q)}{((n + q)(n + 1 - q) + q(k - q))} \\ &= \frac{(n_i + 1)(n + 1 - k)}{((n + k)(n + 1 - k) + k(k - k))} \\ &= \frac{(n_i + 1)}{(n + k)}, \text{ Laplace's Law of Succession} \end{aligned}$$

Natural Law of Succession

Based on uniform cardinality prior:

$$p_C(i | \{n_i\}, n) =$$

$$\begin{cases} (n_i + 1)/(n + k), & \text{if } q = k \\ (n_i + 1)(n + 1 - q)/(n^2 + n + 2q) & \text{if } q < k \cap n_i > 0 \\ q(q + 1)/(k - q)(n^2 + n + 2q) & \text{otherwise} \end{cases}$$

Novel symbol probabilities

$$\text{Let } p(Q | \{n_i\}, n) \doteq \sum_{\{i:n_i>0\}} p(i | \{n_i\}, n)$$

$$\text{Let } p(\bar{Q} | \{n_i\}, n) \doteq 1 - p(Q | \{n_i\}, n)$$

$$p_L(\bar{Q} | \{n_i\}, n) = \frac{k - q}{n + k}$$

$$p_\lambda(\bar{Q} | \{n_i\}, n) = \frac{(k - q)\lambda}{n + k\lambda}$$

$$p_S(\bar{Q} | \{n_i\}, n) = \frac{q(k - q)}{((n + q)(n + 1 - q) + q(k - q))}$$

$$p_C(\bar{Q} | \{n_i\}, n) = \frac{q(q + 1)}{n^2 + n + 2q}$$

Novel symbols and the RNG

After $2^{32}-1$ trials with distinct results, what is the probability of observing the last symbol?

$$p_{\lambda}(\bar{Q} | \{n_i\}, n) = \frac{(k - q)\lambda}{n + k\lambda} = \frac{\lambda}{((1 + \lambda)2^{32} - 1)}$$

$$p_C(\bar{Q} | \{n_i\}, n) = \frac{q(q + 1)}{n^2 + n + 2q} = \frac{2^{32}}{(2^{32} + 2)}$$

Experimental Results

We compare our 2 succession laws with 4 common multinomial estimators from the text compression community.

<u>Method</u>	<u>$p(i \{n_i\}, n)$</u>	<u>$p(\bar{Q} \{n_i\}, n)$</u>
A	$n_i / (n + 1)$	$1 / (n + 1)$
B	$(n_i - 1) / n$	$(k - q)q / (k - r)n$
C	$n_i / (n + q)$	$q / (n + q)$
D	$(n_i - .5) / n$	$q / 2n$

*Note: $r \equiv |\{i: n_i > 1\}|$

Sunrise prediction

Data: We believe the sun has risen for the past 5,220 years without fail.

Plot: days vs. bits

- days that another sunrise is observed

- bits \equiv negative log of total probability that the sun will rise n consecutive days

A smaller slope implies greater confidence in the next sunrise.

First 100 days of history

Best to worst

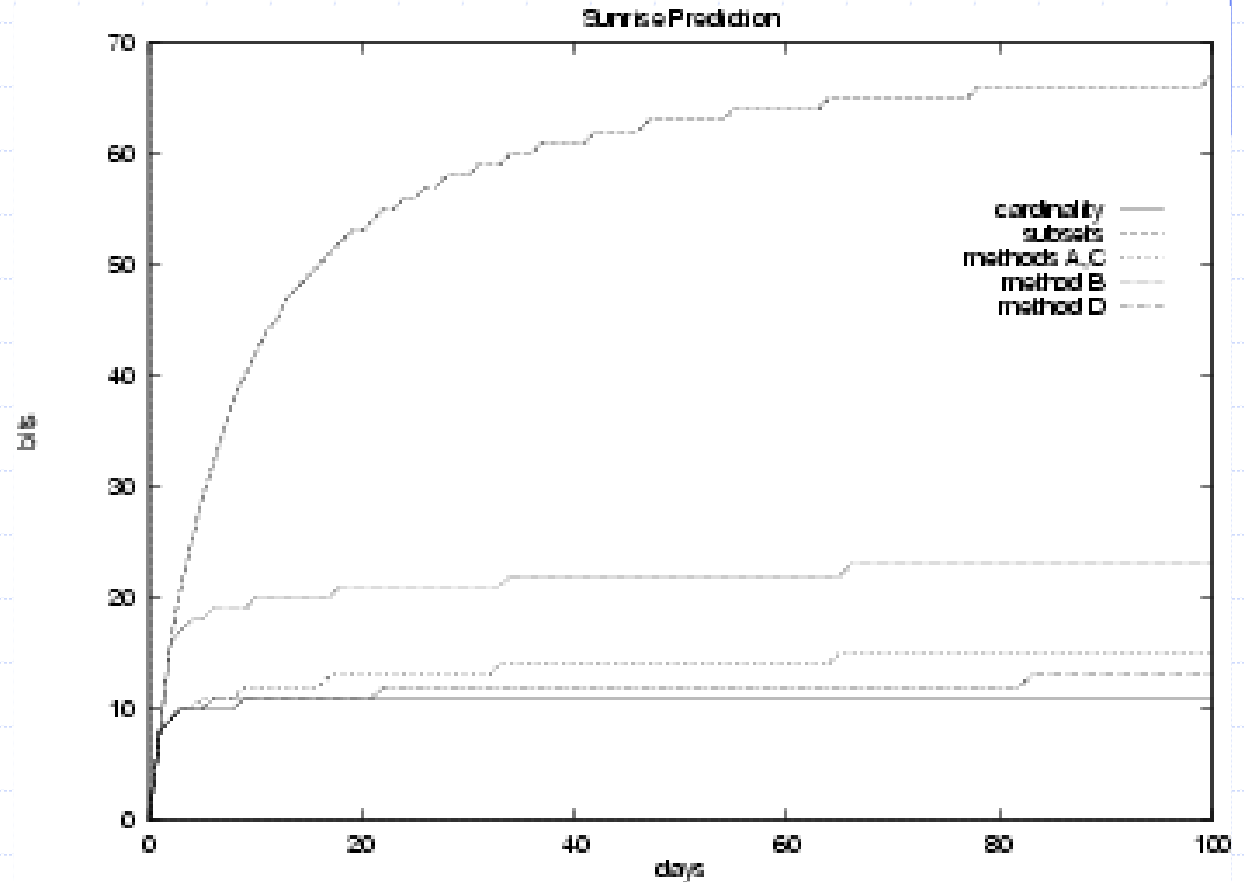
1. Cardinality

2. D

3. A, C

4. B

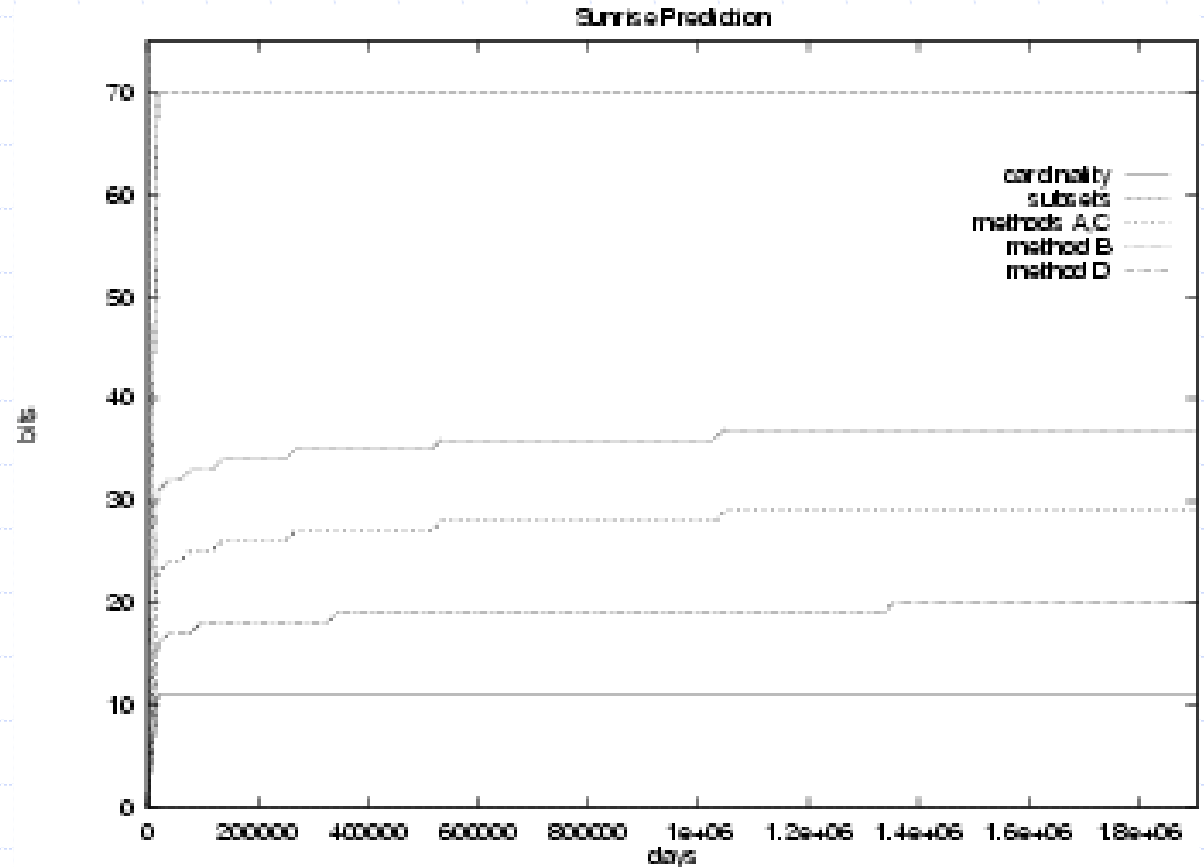
5. subsets



5,220 years of history

Flattest

1. cardinality
2. subsets



A-D do not reach the flatness cardinality had at day 9.

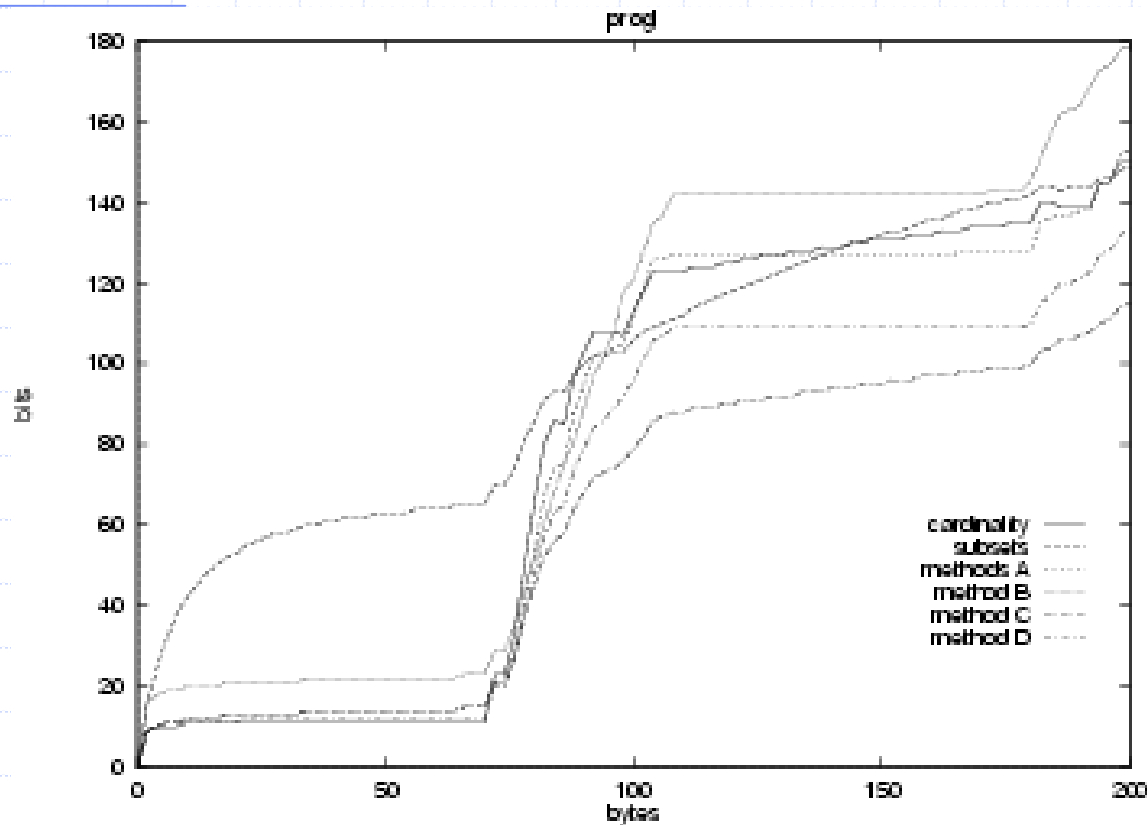
Calgary Corpus

This test set contains both ASCII and non-ASCII files.

Cardinality performed best for almost all files with $q < 100$.

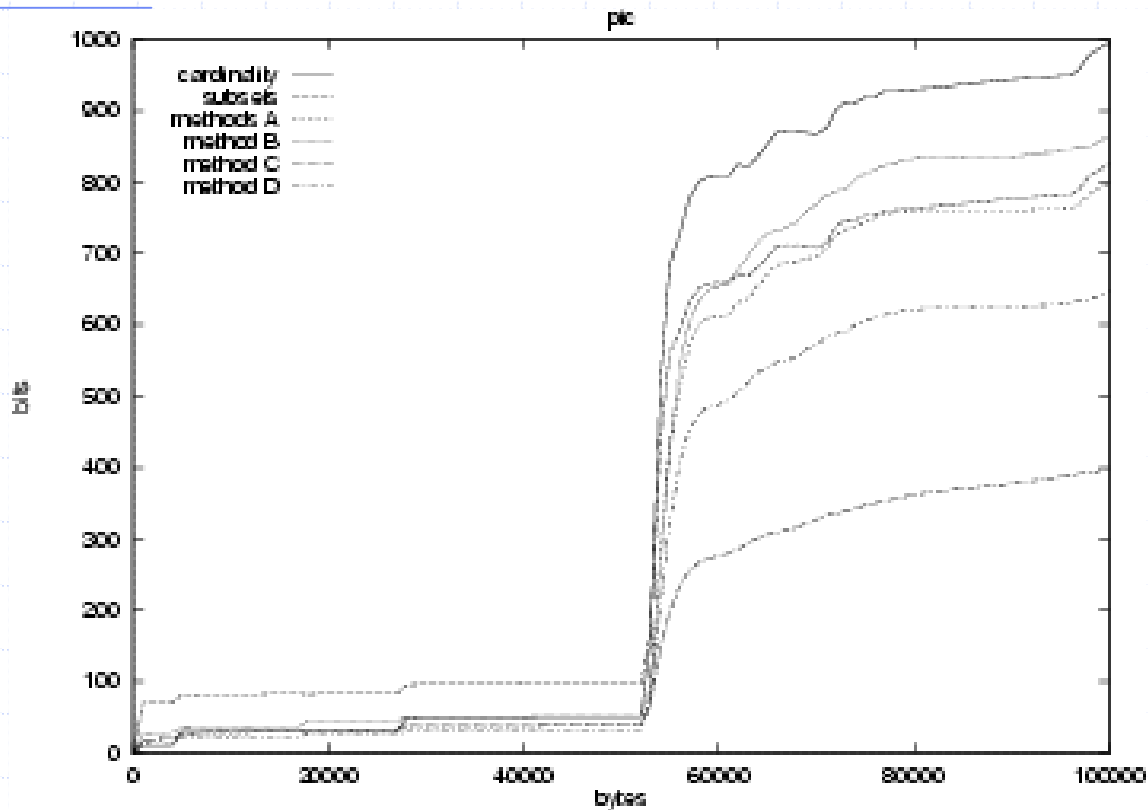
The exception was a LISP file. 71 semi-colons led the cardinality method to predict incorrectly.

Results for progl



Cardinality method converged quickly to multiple ‘;’
but not ideally in this case.

Monochrome picture



The first 52,422 bytes attest only 3 distinct characters of 159 in image.

Cardinality “looses” to all methods except Laplace.

Natural language corpora

We tested prediction methods with the King James Bible and Brown Corpus.

The uniform cardinality law provides best predictions on both counts.

<u>File</u>	q	$p_{\underline{C}}(Q)$	$p_{\underline{S}}(Q)$	$p_{\underline{L}}(Q)$	$p_{\underline{.5}}(Q)$	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>
KJV	74	116	124	427	261	123	221	197	154
Brown	90	141	146	425	265	145	253	226	117

Conclusions

Naturally occurring strings are usually finite and use a small subset of an entire alphabet.

The natural succession law performs best in many areas.

